affyGG

Alberts, Rudi; Vera, Gonzalo; Jansen, Ritsert C.

Published in:
Bioinformatics

DOI:
10.1093/bioinformatics/btm614

Publication date:
2008

BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 3 2008, pages 433–434
doi:10.1093/bioinformatics/btm614

Gene expression

affyGG: computational protocols for genetical genomics with Affymetrix arrays

Rudi Alberts, Gonzalo Vera and Ritsert C. Jansen*

Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Haren, The Netherlands

Received on July 4, 2007; revised on November 22, 2007; accepted on December 10, 2007

Advance Access publication December 16, 2007

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Affymetrix arrays use multiple probes per gene to measure mRNA abundances. Standard software takes averages over probes. Important information may be lost if polymorphisms in the mRNA affect the hybridization of individual probes.

Results: We present custom software to analyze genetical genomics experiments in human, mouse and other organisms: (i) an R package providing functions for QTL analysis at the individual probe level and (ii) Perl scripts providing custom tracks in the UCSC Genome Browser to check for sequence polymorphisms in probe regions.

Availability: http://gbic.biol.rug.nl/supplementary

Contact: r.c.jansen@rug.nl

1 INTRODUCTION

In genetical genomics (Jansen and Nap, 2001), gene expression profiles of genetically different individuals are combined with molecular markers in the DNA to reveal expression quantitative trait loci (QTL). Especially the Affymetrix technology, which is a popular platform for profiling gene expression, poses many challenges in data analysis since it uses multiple 25 mer probes per gene to measure its mRNA abundance. Alberts *et al.* (2005) proposed a statistical multi-probe model: $\log(y_{ij}) = m + P_j + G_i + PG_{ij} + e_i + e_{ij}$, where y_{ij} is the signal of the jth probe of the ith individual, P_j is the probe effect and G_i is the genotype effect at the marker under study. PG_{ij} is the probe-specific genotype effect, which may be caused by (unassayed) single nucleotide polymorphisms (SNPs) in probe regions leading to a difference in hybridization, and not in gene expression (Alberts *et al.*, 2007a). The model is computed at all marker positions to find the position (QTL) with the most significant G_i; the significance of the corresponding PG_{ij} quantifies the probe specificity of the QTL (QTL × probe). Using parametric bootstrap the significance is calculated from data drawn from the no-QTL model $\log(y_{ij}) = m + e_i + e_{ij}$. See Alberts *et al.* (2005) and Alberts *et al.* (2007a) for applications in human (association analysis of cell lines) and mouse (linkage analysis of recombinant inbred lines). The model is also applicable to backcross, intercross and other experimental designs (optionally with additive QTL effects). The model can also include a batch factor to remove unwanted batch effects.

*To whom correspondence should be addressed.

2 PROTOCOL

Figure 1 shows the workflow from the raw Affymetrix data in the form of .CEL files to QTL visualization. In the pre-processing part, the .CEL files are background corrected and normalized using the RMA method (Irizarry *et al.*, 2003). In the processing part, QTL analysis is performed as described in Alberts *et al.* (2005), and deviating probes are detected using a procedure for backward elimination as described in Alberts *et al.* (2007a). Finally, a custom track in the UCSC Genome Browser is created, visualizing individual Affymetrix probes and known sequence polymorphisms (SNPs, splicing variants) in probe regions. The affyGG software accepts missing marker genotypes, but excludes them from the QTL analysis. In crosses, most likely genotypes can be imputed in advance using R/QTL (www.rqtl.org/manual/html/argmax.geno.html).

2.1 Pre-processing

(1) Create and load in R a comma separated values (CSV) file containing the genotype data (e.g. genotypes.csv), with format:
{molecular marker names} x {individuals}, where the cells contain the genotype labels (e.g. 1 or 2; AA, AC or CC; U for missing data).
> genotypes ← read.csv('genotypes.csv')

(2) Create and load in R a CSV file (e.g. markerpositions.csv) containing the positions of the markers, with format:
{molecular marker names} x {chromosome number (chr), position in basepairs (bp)}
> markersPos ← read.csv('markerpositions.csv')

(3) Create a vector containing the names of the .CEL files to be used, and specify the directory where the .CEL files are located:
> celfiles ← c(''bxd1a.cel'', ''bxd2a.cel'', ''bxd5a.cel'', ''bxd6a.cel'', . . .)
> directory ← ''C: /myproject/celfiles''

(4) Run the pre-processing function by typing:
> probesignals ← rma.preprocessing(directory, celfiles, *filename* = ''probesignals.csv'')

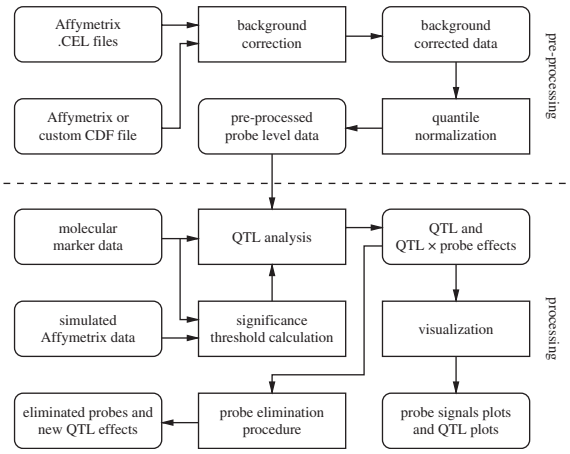

Fig. 1. Pre-processing and processing of Affymetrix array data. Squared boxes are R functions, rounded boxes are data.

Custom CDF files, as provided by Alberts *et al*. (2007b), can be used as follows:
>probesignals ← rma.preprocessing(directory, celfiles, *filename* = "probesignals.csv", *cdffile* = cdffile, *cdfpackage* = cdfpackage)

2.2 QTL analysis

(5) Specify in which batch each individual was processed:
>batch ← c(2, 2, 2, 1, 1, 2, 1, 1, 2, 3, 2, 2, 2, 2, 3, 3, 3, 3, 2, 3, 3, 1, 1, 1, 2, 3, 1, 2, 1, 1)

(6) Select the probe level data for one probe set:
>traits ← probesignals[probesignals$probeset = = '96254_f_at',]

(7) Perform QTL analysis on probe level:
>qtlmap ← qtlMap.xProbe(*genotypes* = genotypes, *traits* = traits, *batch* = batch)
>qtlProbe ← cbind(qtlmap[,1:3], minlog10pmarker = −log10(qtlmap[,4]))

(8) Perform QTL analysis on probe set level:
>qtlmapProbeset ← qtlMap.xProbeSet(*genotypes* = genotypes, *traits* = traits, *batch* = batch)
>qtlProbeset ← −log10(qtlmapProbeset$pmarker)
>intProbeset ← −log10(qtlmapProbeset$pinteraction)

(9) Calculate the genome-wide significance threshold:
>qtlThres ← qtlThresholds.sma(*genotypes* = genotypes, *batch* = batch, *nProbes* = 16, *nIndiv* = 30, *n.simulations* = 1000, *filename* = "qtlThres.csv")

2.3 Visualizations

(10) Create plot of the probe signals for a given probe set (Figure 2):
Specify the interrogation positions of the probes on the mRNA:
>pos ← c(1893, 1894, 1897, 1904, 1906, 1911, 1912, 1913, 1916, 1925, 1929, . . .)
>probePlot(*traits* = traits, *probesetName* = "96243_f_at", *probesPos* = pos, *alleleColors* = mycolors)

(11) Create QTL plots for each of the probes of one probe set:
Collect the starting positions of each chromosome (in basepairs):
>chrOffsets ← c(0.000000, 197.842934, 379.178330, 540.742912, 693.473822, . . .)
>qtlPlot.xProbe(*probesetName* = "96243_f_at", *markersPos* = markersPos, *probeQtlProfiles* = qtlProbe, *qtlThres* = 3.72, *chrOffsets* = chrOffsets, *filename* = "out1.png")

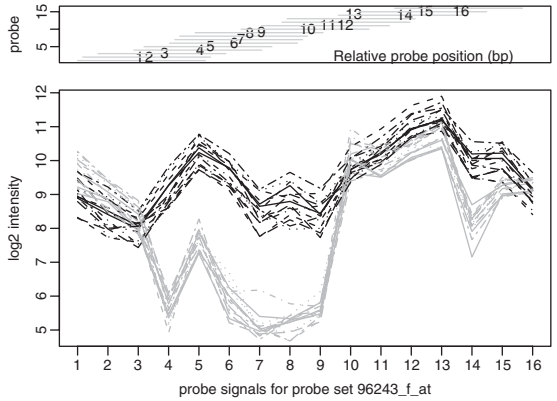

Fig. 2. Affymetrix probe level data for 30 mice. The mice with black (grey) profile carry the B6 (D2) allele at marker D15mit158. The probes have been designed for B6. Two unassayed and previously unknown SNPs in D2 (one in probes 4–9, the other in probes 11–15) explain the lower signals of the grey profiles.

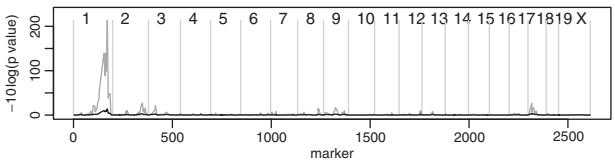

Fig. 3. QTL plot on probe set 96243_f_at. The black (gray) curve corresponds to the significance of QTL (QTL × probe).

(12) Create QTL plots on probe set level (Figure 3):
>qtlPlot.xProbeSet(*probesetName* = "96243_f_at", *markersPos* = markersPos, *probesetQtlProfile* = qtlProbeset, *interactionProfile* = intProbeset, *qtlThres* = 3.65, *interactionThres* = 3.75, *chrOffsets* = chrOffsets, *filename* = "out2.png")

2.4 Eliminate and check deviating probes

(13) Eliminate deviating probes using the statistical method developed in Alberts *et al*. (2007a):
>pe ← probeElimination(*probesetName* = "160371_at", *markerName* = "D7Mit100", *genotypes* = genotypes, *traits* = traits, *batch* = batch)

(14) Users can download our Perl scripts to check probes in the UCSC Genome Browser (http://genome.ucsc.edu) for known SNPs not used in the genotyping of the mapping population.

Conflict of Interest: none declared.

REFERENCES

Jansen,R.C. and Nap,J.P. (2001) Genetical genomics: the added value from segregation. *Trends Genet*., **17**, 388–391.

Irizarry,R.A. *et al*. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Alberts,R. *et al*. (2005) A statistical multiprobe model for analyzing cis and trans genes in genetical genomics experiments with short-oligonucleotide arrays. *Genetics*, **171**, 1437–1439.

Alberts,R. *et al*. (2007a) Sequence polymorphisms cause many false cis eQTLs. *PLoS ONE*, **2**, e262.

Alberts,R. *et al*. (2007b) A verification protocol for the probe sequences of Affymetrix genome arrays reveals high probe accuracy for studies in mouse, human and rat. *BMC Bioinformatics*, **8**, 132.